Experimentation in Software Engineering – edition 2024

TRAINING ANSWERS

This chapter contains answers to the questions in Section A.1. Some of the answers should not be seen as absolute but instead as suggestions. To solve the statistical problems StatView[1] 5.0 has been used.

1 *Normally distributed data*

The results from increasing the number of segments from ten to twelve can be found in Table 1. The boundaries were calculated with the help of tables in [Humphrey95]. The values from the tables were multiplied by the standard deviation, and then the mean was added. All data was cross-checked by calculating the boundary values by hand.

Table 1. Segments

Segment	Lower Boundary	Upper Boundary	Number of values
1	$-\infty$	678.8	5
2	678.8	713.8	7
3	713.8	738.3	3
4	738.3	758.8	6
5	758.8	777.3	5
6	777.3	795.0	3
7	795.0	812.6	7
8	812.6	831.2	3
9	831.2	851.7	5
10	851.7	876.2	6
11	876.2	911.1	6
12	911.1	∞	4

[1] The development of StatView was discontinued, and the application is no longer available.

The excepted number of values in each segment is 5. This means that $X^2 = 4.8$. The number of degrees of freedom is 12–21 = 9. In Table A2, it can be seen that $\chi^2_{0.05,9} = 16.92$. Since $X^2 < \chi^2_{0.05,9}$ it is impossible to reject the null hypothesis at the 0.05 level.

Another possibility is to use the predefined template in StatView to test for normality. It automatically calculates the ideal normal values for the data, and the two data sets are compared using a Kolmogorov-Smirnov test. The results from this test can be found below.

Kolmogorov-Smirnov Test for Measurement
Grouping Variable: Category for Measurement

DF	2
Count, Actual	60
Count, Ideal Normal	60
Maximum Difference	.050
Chi Square	.300
P-Value	>.9999

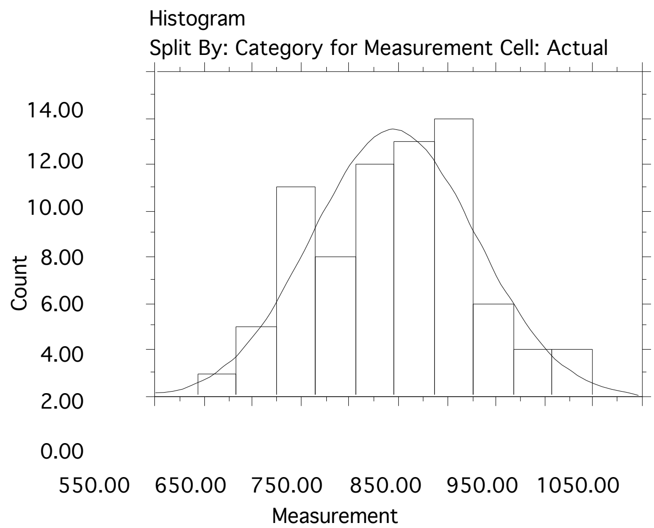

Figure 1. Normality test results.

2 *Experience*

2.1 Question 1

To extend the survey, different ratio measures could be added, or some of them could be transformed into ratio scales, such as years of programming knowledge. Other interesting issues could be:

- Age
- Gender
- Experience of the development environment
- Experience of the development platform
- Mathematical knowledge
- Grading of activities according to how interesting and fun they are, for example, requirements specification, design, programming, and quality assurance (inspections and testing).

2.2 Question 2

It is possible to construct several different hypotheses and many of them would probably provide significant results. The problem is that the hypothesis might be of small or non-existent relevancy. Other possible hypotheses can be:

1. H_0: There is no difference in terms of relative size prediction error and general knowledge in computer science and software engineering.
   H_1: The relative size prediction error changes with general knowledge in computer science and software engineering.
   Knowledge in computer science and software engineering can affect the prediction accuracy since they are aware of metrics collection and understand the concepts of size estimation. A person also might have better control of the personal development process.
2. H_0: There is no difference between productivity and general programming knowledge.
   H_1: The productivity changes with general programming knowledge.
   If general programming knowledge influences the productivity, then perhaps it is not necessary with knowledge in any specific programming language.
3. H_0: The number of faults does not affect development time.
   H_1: The number of faults does affect development time.
   The intention is to investigate if it is possible to have many faults without increasing the development time. If that is true, an investigation that includes defect types could be the next step.

2.3 Question 3

The sampling is a non-probability convenience sampling, where the most convenient persons are all the people attending the course.

2.4 Question 4

All the hypotheses except the last one were tested with factorial ANOVA, i.e. assigning one or more nominal variables (factors) to one or more dependent variables. Hypothesis 3 was instead tested with Pearson correlation. The results from the tests can be found below, along with an interpretation.

Hypothesis 1.

ANOVA Table for Pred. size

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
S.E. or C.S.	3	1261,313	420,438	1,911	,1385	5,733	,458
Residual	55	12099,914	219,998				

Figure 2. ANOVA results for hypothesis 1.

The result was not significantly different because the P-value was 0.1385.

Hypothesis 2.

ANOVA Table for Prod.

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Programming	3	580.042	193.347	3.080	.0349	9.239	.688
Residual	55	3453.048	62.783				

Figure 3. ANOVA results for hypothesis 2.

The result was significantly different because the P-value was 0.0349.

Hypothesis 3.

Correlation Analysis

	Correlation	P-Value	95% Lower	95% Upper
Time, Faults	,475	,0001	,250	,652

59 observations were used in this computation.

Figure 4. ANOVA results for hypothesis 3.

In this case, the correlation calculation was significant, but the correlation itself was too small. It should have been at least 0.75, but this value depends on the purpose of the study. Also, it might be a non-linear relationship.

2.5 Question 5

Two of the hypotheses, except hypothesis 1, could be rejected because they had low significance except for the correlation. The problem with that one was that the correlation was too small. However, we could reject the null hypothesis that there is no difference between general knowledge in programming and productivity.

If the results had been different, there still should have been some problems with the external validity. The subjects are students, and they work in a special environment where they can easily ask a friend for help and get a quick response. Also, knowledge about software engineering and computer science is mostly on an academic level that can differ from that of the industry. The programs created are small, and the problems are fairly easy to solve compared to the large complex systems in industry.

3 Programming

3.1 Question 1

This study should be divided into four separate studies where each of them has a completely randomized and balanced design.

3.2 Question 2

The hypotheses can be defined as follows:

1. H_0: There is no difference in terms of program size between persons using programming languages A and B.
   H_1: There is a difference in terms of program size according to programming language.
2. H_0: There is no difference in terms of development time between persons using programming languages A and B.
   H_1: There is a difference in terms of development time according to programming language.
3. H_0: There is no difference in terms of the number of defects between persons using programming languages A and B.
   H_1: There is a difference in terms of the number of defects according to programming language.
4. H_0: There is no difference in terms of defects found in the test between persons using programming languages A and B.
   H_1: There is a difference in terms of defects found in the test according to programming language.

3.3 Question 3

In all the cases there are two outliers, one that has a higher value and one that has a smaller value. Since they appear in all the plots and on both sides of the median it seems to be a very consequent outliers and should therefore probably not be removed.

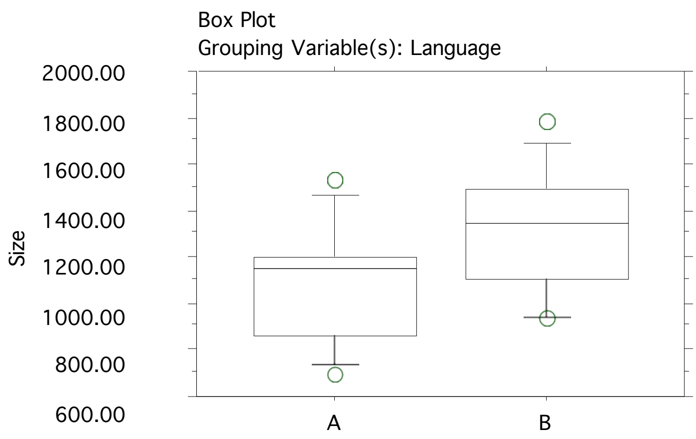

Figure 5. Box-plot of the size versus the language

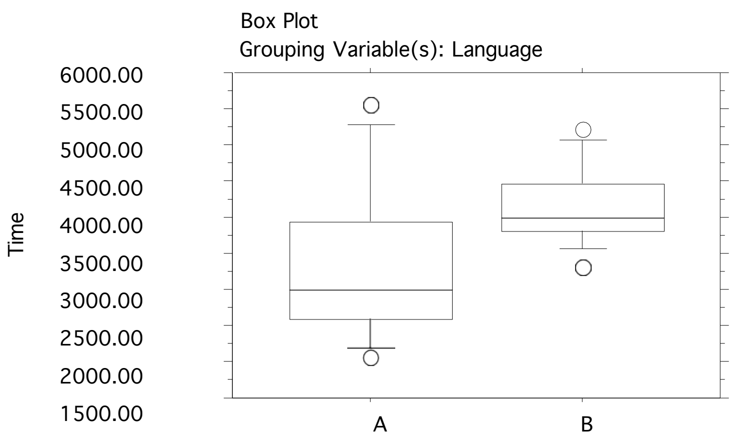

Figure 6. Box-plot of the time versus the language.

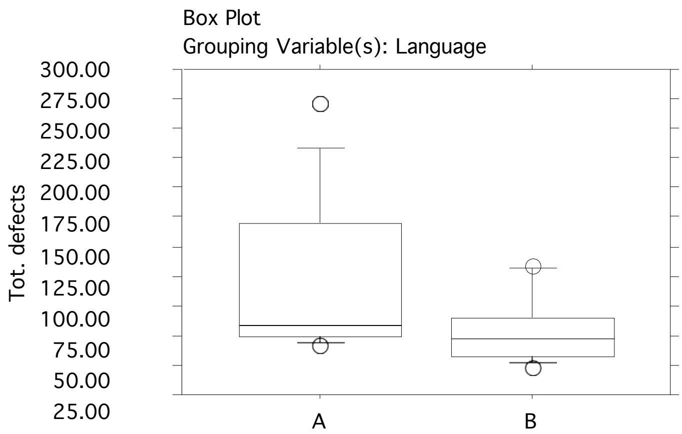

Figure 7. Box-plot of the total number of defects versus the language.

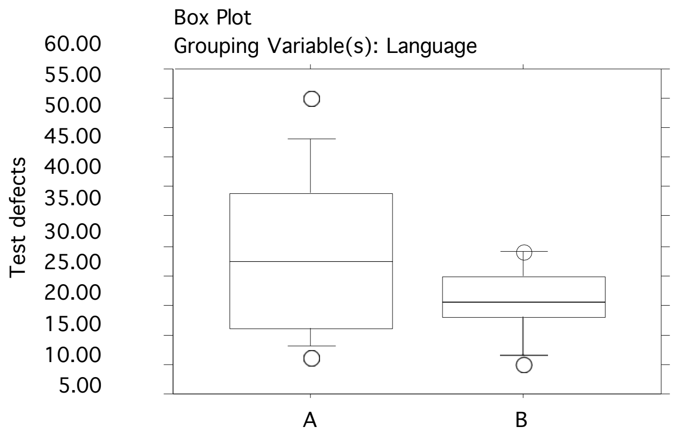

Figure 8. Box-plot of the number of test defects versus the language.

3.4 Question 4

The intention is to use a parametric comparison, and the most appropriate analysis method is, therefore, an unpaired t-test. The results can be found below.

Hypothesis 1.

Unpaired t-test for Size
Grouping Variable: Language
Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
A, B	-227,500	18	-1,867	,0782

Group Info for Size
Grouping Variable: Language

	Count	Mean	Variance	Std. Dev.	Std. Err
A	10	1092,700	72777,344	269,773	85,310
B	10	1320,200	75653,956	275,053	86,979

Figure 9. Results of a t-test for hypothesis 1.

It is possible to reject H_0 at the 0.1 level. Subjects using language B create a significantly larger program.

Hypothesis 2.

Unpaired t-test for Time Grouping Variable: Language Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
A, B	-779,100	18	-1,892	,0746

Group Info for Time Grouping Variable: Language

	Count	Mean	Variance	Std. Dev.	Std. Err
A	10	3402,900	1371420,989	1171,077	370,327
B	10	4182,000	323365,333	568,652	179,824

Figure 10. Results of a t-test for hypothesis 2.

It is possible to reject H_0 at the 0.1 level. Subjects using language B spend more effort developing their program.

Hypothesis 3.

Unpaired t-test for Tot. defects
Grouping Variable: Language Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
A, B	39,200	18	1,647	,1170

Group Info for Tot. defects
Grouping Variable: Language

	Count	Mean	Variance	Std. Dev.	Std. Err
A	10	119,400	4776,489	69,112	21,855
B	10	80,200	891,067	29,851	9,440

Figure 11. Results of a t-test for hypothesis 3.

It is not possible to reject H_0 at the 0.1 level even though it was very close. Therefore, there is no significant difference between the total number of faults introduced using language A or B.

Hypothesis 4.

Unpaired t-test for Test defects
Grouping Variable: Language Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
A, B	7,600	18	1,616	,1234

Group Info for Test defects Grouping Variable: Language

	Count	Mean	Variance	Std. Dev.	Std. Err
A	10	28,400	182,933	13,525	4,277
B	10	20,800	38,178	6,179	1,954

Figure 12. Results of a t-test for hypothesis 4.

It is not possible to reject H_0 at the 0.1 level. There is no significant difference between the defects found in test using language A or B.

3.5 Question 5

The intention is to use a non-parametric comparison and the most appropriate analysis method is therefore a Mann-Whitney test. The results can be found below. Compared to the t-tests, the results from the non-parametric Mann-Whitney tests are a little bit more restrained. This is true for all the hypotheses except number two even though the overall difference is not large.

Hypothesis 1.

Mann-Whitney U for Size
Grouping Variable: Language

U	28,000
U Prime	72,000
Z-Value	-1,663
P-Value	,0963
Tied Z-Value	-1,663
Tied P-Value	,0963
# Ties	0

Figure 13. **Mann-Whitney test results for hypothesis 1.**

It is possible to reject H_0 at the 0.1 level.

Hypothesis 2.

Mann-Whitney U for Time
Grouping Variable: Language

U	26,000
U Prime	74,000
Z-Value	-1,814
P-Value	,0696
Tied Z-Value	-1,814
Tied P-Value	,0696
# Ties	0

Figure 14. **Mann-Whitney test results for hypothesis 2.**

It is possible to reject H_0 at the 0.1 level.

Hypothesis 3.

Mann-Whitney U for Tot. defects
Grouping Variable: Language

U	30,000
U Prime	70,000
Z-Value	-1,512
P-Value	,1306
Tied Z-Value	-1,513
Tied P-Value	,1303
# Ties	2

Figure 15. **Mann-Whitney test results for hypothesis 3.**

It is not possible to reject H_0 at the 0.1 level.

Hypothesis 4.

Mann-Whitney U for Test defects
Grouping Variable: Language

U	33,500
U Prime	66,500
Z-Value	-1,247
P-Value	,2123
Tied Z-Value	-1,250
Tied P-Value	,2113
# Ties	3

Figure 16. **Mann-Whitney test results for hypothesis 4.**

It is not possible to reject H_0 at the 0.1 level.

3.6 Question 6

In this case, where our measures are on a ratio scale, it is appropriate to use a parametric test. As we can see, the two different methods provided almost the same results. Noticeable is that the last hypothesis could not be rejected in any of the tests, but in the Mann-Whitney test, the results were not even close to 0.1, so we should be tempted not to reject it.

The advantage of the non-parametric analysis is that the approach is a little bit more careful, i.e. we are more secure about our statements. The disadvantage is that we might not reject H_0 even though there is a significant difference.

3.7 Question 7

The problem is that in this study, we do not know anything about the subjects' previous knowledge of the programming language. This might affect the results. If the subjects had been able to choose the programming language themselves, the study would not have been valid, nor would the results. The problem is that the subjects perhaps choose a specific language because they want to learn that language or they might already have knowledge about it. A solution could be to collect data about the subjects and change the design and block out unwanted factors.

4 Design

4.1 Question 1

The intention with this study is to evaluate the impact of quality object-oriented design principles modifying a given design. In this case the object is the design document, and the treatments should be seen as the design principles because it is impossible to separate between the different object types. Thus, the experiment has a paired design.

4.2 Question 2

The hypotheses can be defined as follows:

1. H_0: There is no difference in terms of time spent on identifying places for modification between the good and bad quality design documents.
   H_1: There is a difference in terms of time spent on identifying places for modification between the good and bad quality design documents.
2. H_0: There is no difference in terms of completeness of the impact analysis between the good and bad quality design documents.
   H_1: There is a difference in terms of completeness of the impact analysis between the good and bad quality design documents.
3. H_0: There is no difference in terms of the correctness of the impact analysis between the good and bad quality design documents.
   H_1: There is a difference in terms of the correctness of the impact analysis between the good and bad quality design documents.
4. H_0: There is no difference in terms of modification rate between the good and bad quality design documents.
   H_1: There is a difference in terms of modification rate between the good and bad quality design documents.

4.3 Question 3

The subjects that are missing values should not be included in those analysis parts for which they are missing values. The problem is that we will have fewer data points, which will make it more difficult to get significant results.

4.4 Question 4

The intention is to use a parametric comparison and the most appropriate analysis method is therefore a paired t-test. The results can be found below.

Hypothesis 1.

Paired t-test
Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
Time (good), Time (bad)	-,588	16	-,251	,8048

Descriptive Statistics

	Mean	Std. Dev.	Std. Error	Count	Minimum	Maximum	# Missing
Time (good)	29,714	13,413	2,927	21	9,000	65,000	12
Time (bad)	30,211	9,739	2,234	19	10,000	50,000	14

Figure 17. **Results of a paired t-test for hypothesis 1.**

In this case there is almost no difference, and we cannot reject H_0. The subjects spent equally much effort on finding correct places in both the good and the bad design.

Hypothesis 2.

Paired t-test
Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
Completeness (good), Completeness (bad)	,207	31	4,176	,0002

Descriptive Statistics

	Mean	Std. Dev.	Std. Error	Count	Minimum	Maximum	# Missing
Completeness (good)	,663	,269	,047	33	0,000	1,000	0
Completeness (bad)	,452	,219	,039	32	,095	,762	1

Figure 18. **Results of paired t-test for hypothesis 2.**

There is a significant difference at the 0.001 level according to the number of correct places found, i.e. the subjects found more of the correct places in the good design.

Hypothesis 3.

Paired t-test
Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
Correctness (good), Correctness (bad)	,023	29	,940	,3548

Descriptive Statistics

	Mean	Std. Dev	Std Error	Count	Minimum	Maximum	# Missing
Correctness (good)	,950	,183	,033	31	0,000	1,000	2
Correctness (bad)	,930	,125	,022	32	,500	1,000	1

Figure 19. **Results of a paired t-test for hypothesis 3.**

There is no significant difference in terms of number of correct places found according to number of places indicated as found. This means that equally large number of false positives indicated in both the good and the bad design.

Hypothesis 4.

Paired t-test
Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
Modification (good), Modification (bad)	,213	16	2,393	,0293

Descriptive Statistics

	Mean	Std. Dev	Std Error	Count	Minimum	Maximum	# Missing
Modification (good)	,645	,401	,087	21	0,000	1,692	12
Modification (bad)	,375	,309	,071	19	,080	1,400	14

Figure 20. **Results of paired t-test for hypothesis 4.**

There is a significant difference on the 0.05 level that it is more efficient to find correct places in the good design than in the bad.

4.5 Question 5

It is possible to use two different non-parametric analysis methods, the Paired sign test and the Wilcoxon signed rank test. We have chosen the latter one and the results can be found below. We have been able to reject the same H_0 hypotheses as with the parametric analysis methods.

Hypothesis 1.

Wilcoxon Signed Rank Test for Time (good), Time (bad)

# 0 Differences	4
# Ties	2
Z-Value	-,699
P-Value	,4846
Tied Z-Value	-,701
Tied P-Value	,4832

16 cases were omitted due to missing values.

Figure 21. Wilcoxon results for hypothesis 1.

It is not possible to reject H_0.

Hypothesis 2.

Wilcoxon Signed Rank Test for Completeness (good), Completeness (bad)

# 0 Differences	0
# Ties	2
Z-Value	-3,422
P-Value	,0006
Tied Z-Value	-3,422
Tied P-Value	,0006

One case was omitted due to missing values.

Figure 22. Wilcoxon results for hypothesis 2.

It is possible to reject H_0 at the 0.001 level.

Hypothesis 3.

Wilcoxon Signed Rank Test for Correctness (good), Correctness (bad)

# 0 Differences	18
# Ties	2
Z-Value	-1,490
P-Value	,1361
Tied Z-Value	-1,492
Tied P-Value	,1358

Three cases were omitted due to missing values.

Figure 23. **Wilcoxon results for hypothesis 3.**

It is not possible to reject H_0.

Hypothesis 4.

Wilcoxon Signed Rank Test for Modification (good), Modification (bad)

# 0 Differences	0
# Ties	0
Z-Value	-2,107
P-Value	,0352
Tied Z-Value	-2,107
Tied P-Value	,0352

16 cases were omitted due to missing values.

Figure 24. **Wilcoxon results for hypothesis 4.**

It is possible to reject H_0 at the 0.05 level.

4.6 Question 6

In this case the parametric methods were more distinct in their results if H_0 should be rejected or not. Therefore, these parametric methods could provide better and more helpful information. The difference is that the variation for the non-parametric *p*-values is smaller than for the parametric one.

4.7 Question 7

The population is students at the university with some knowledge about software engineering. This affects the external validity because the students are in a special position where they would like to get as good grades as possible and therefore have another kind of motivation. Furthermore, our knowledge about their previous knowledge in related areas is very sparse. To improve the study, subjects from industry could be included and factors that might influence the results can be blocked out.

5 Inspections

The answers to the following questions are not absolute, i.e. there are other solutions and other aspects to investigate.

5.1 Question 1

The intention is to compare the distributions of the defects found per perspective and see if they differ significantly. This should be done for both the PG document and the ATM document separately with a Chi-2 test. The problem is that the data does not fulfil the rule of thumb that no expected value should be less than five. Therefore, it is necessary to group them in some way. Another possibility is to use the Pearson correlation to investigate differences between different perspectives.

5.2 Question 2

To be able to draw valid conclusions, it is necessary to investigate if the two objects, i.e. the documents, have the same complexity or some other factor that might affect the results. By choosing, for example, the defect detection rate between the documents, it is possible to find this kind of relationship. These relationships can be studied with the help of ANOVA tests.

5.3 Question 3

Definition. The objective of this study is to evaluate the difference between the three perspectives in PBR, i.e., do they find the same defects, or are the same defects found which could reduce resources? Furthermore, one perspective might find more defects or be more effective. The definition of the study can be summarized as:

- Analyze the three PBR perspectives for the purpose of evaluation with respect to their defect-finding ability from the point of view of the researcher in the context of students in the software engineering course.

Context. The next step is the context selection. In this case, we have got a situation that is off-line with students as the subjects. The problem is a real one, and the findings are more specific than general.

Hypothesis formulation. Based on the above formulation, we can define the following hypotheses:

1. H_0: There is no difference between the defects found by the three different perspectives.
   H_1: There is a difference between the defects found by the three different perspectives.
   This hypothesis investigates the need for the different perspectives and if it is useful to have them all.
2. H_0: There is no difference between the two documents in terms of complexity.
   H_1: There is a difference between the two documents in terms of complexity.
   The intention with this analysis is to strengthen the conclusions about the previous hypothesis because if one of the documents is more complex and the subjects must spend more effort finding defects, it might affect the study.

Variables selection. In this case, all the variables are already provided. Some of them are not used in this example but are listed anyway to provide insight into alternative data for deeper investigation of some relationship. The collected variables for the subject are the following[2]:

- Perspective – user, developer, or tester
- Document – ATM or PG
- Time – minutes spend on finding defects
- Defects – number of defects found
- Efficiency – 60*Defects/Time
- Rate – Defects/total defects in the document (ATM 29 and PG 30)

Subjects. The subjects in this study are 30 students taking a course in software engineering. The sampling is a non-probability convenience sampling where the most convenient persons are all the persons attending the course.

Experiment design. To test the hypothesis, a 2*3 factorial design should be chosen. The two factors are perspective and document. The experiment varies the three perspectives over the two documents. Because of the setting with students, a formal experiment could not be conducted. Instead, this is a quasi-experiment. To analyse the first hypothesis, a Chi-2 test should be applied, and a Pearson correlation calculation. The second one should be analysed with an ANOVA test.

Threats. One threat to the conclusion validity is the number of samples in the study, which may reduce the ability to reveal patterns. Threats to the internal validity are the selection of subjects and instrumentation. The subjects take a course in software engineering and have not made an active decision to take part in the study, and the selection is not random. Moreover, the documents used may affect the results. There might be issues that could be considered defects even though they are not, and therefore, there may be an increased number of false positives. The construct validity is affected by the size of the

[2] A more detailed description is provided in the training question.

sample, i.e. there are not enough students for each perspective and document. Finally, the external validity is affected by using students to be able to generalize the results. This is always a problem in these kinds of studies.

Descriptive statistics. The first thing is to visually analyse the distribution of the different defects found. As we can see, many of the defects found have the same detection rate for all of three perspectives. In this case it is not necessary to look for outliers since the counts can only vary between zero and five.

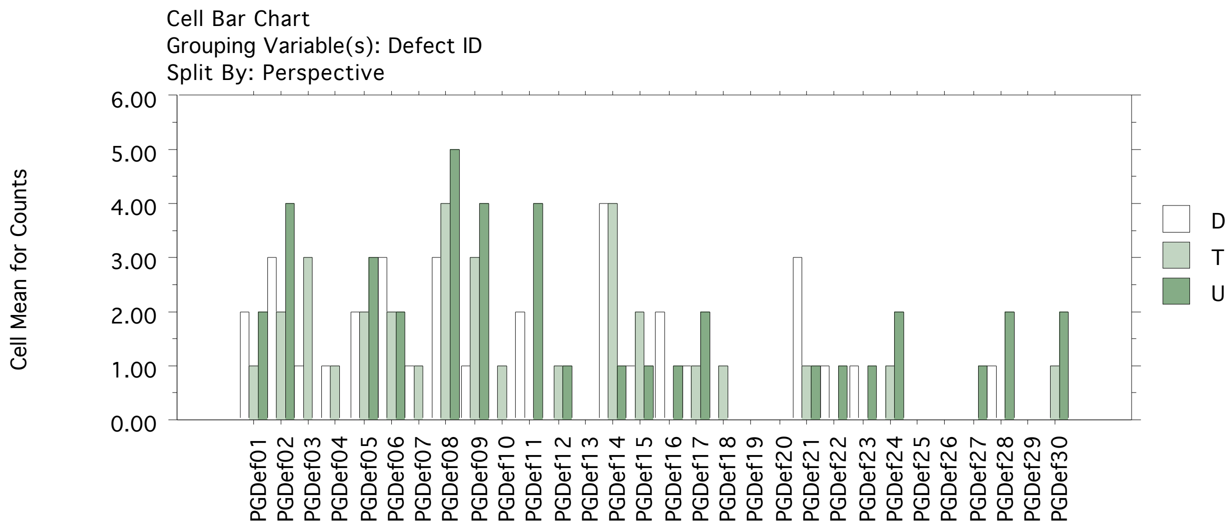

Figure 25. Cell bar chart for PG defects found by different perspectives.

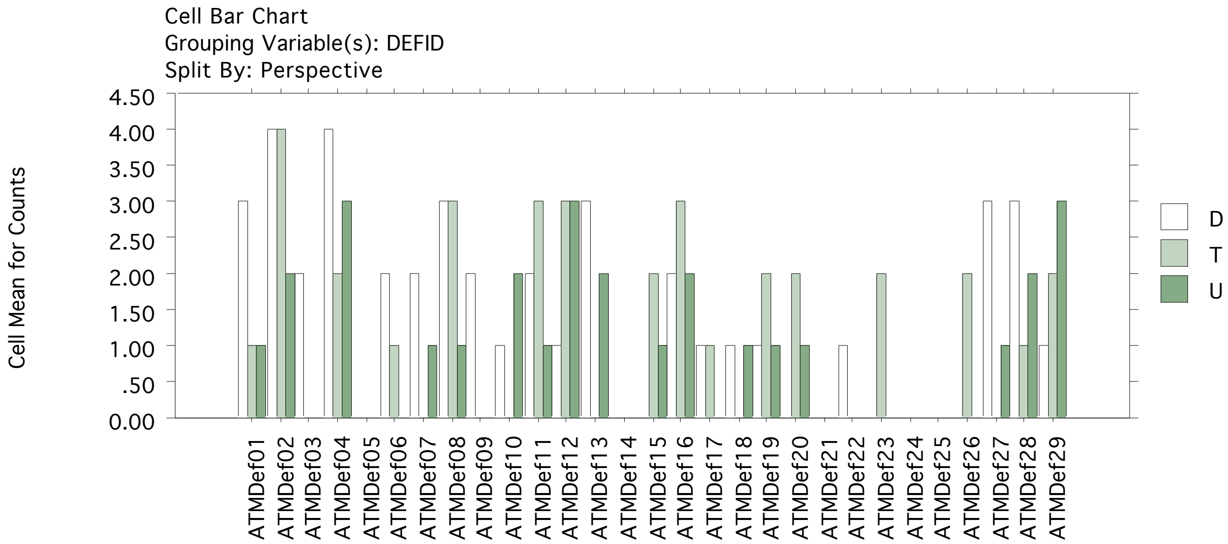

Figure 26. Cell bar chart for ATM defects found by different perspectives.

The boxplots for the time, efficiency and rate variables show variations among the documents with respect to the perspectives. For example, in terms of the effort spent by the reviewers from the user perspective, the subjects reviewing the ATM seem to spend much more effort.

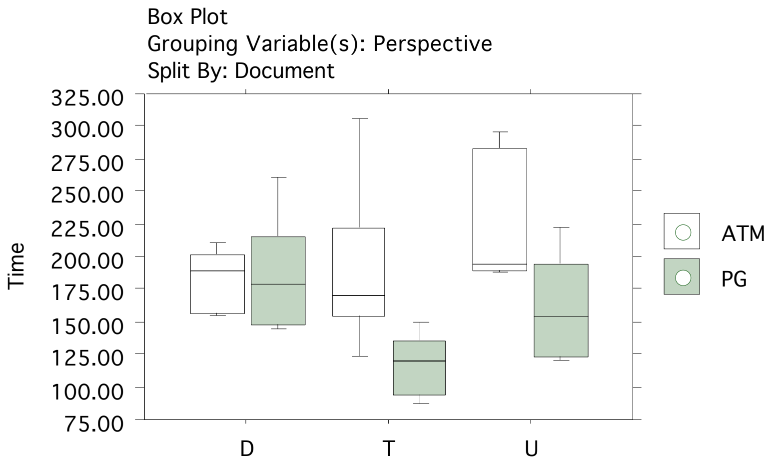

Figure 27. Box-plot for effort spend on reviewing.

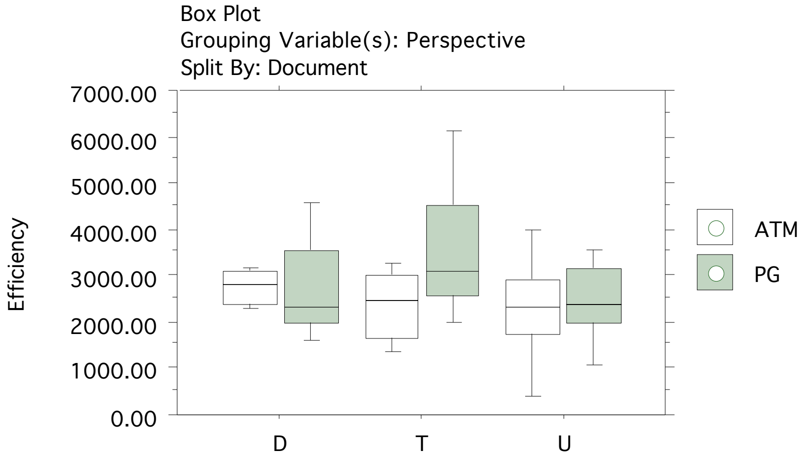

Figure 28. Box-plot of efficiency for the different perspectives and documents.

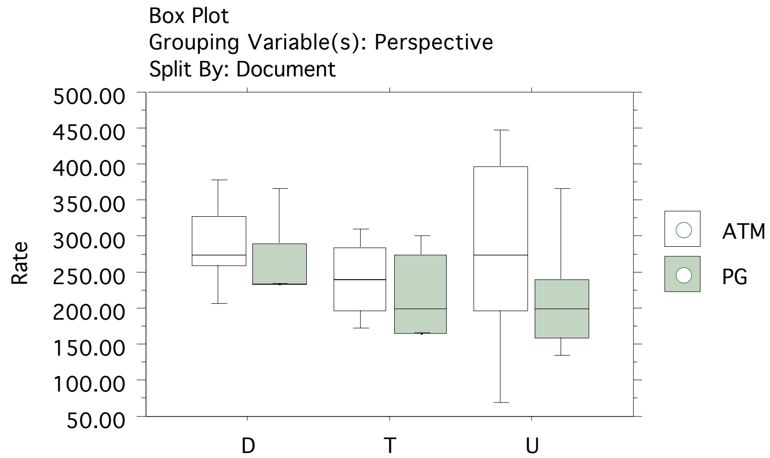

Figure 29. Box-plot of detection rate for the different perspectives and documents.

Data set reduction. As mentioned earlier, the boxplots did not show any potential outliers, so therefore, no data set reduction is necessary.

Hypothesis testing. To test the hypothesis of no difference between the different perspectives, a Chi-2 test was performed. The problem is that the data does not fulfil the rules of thumb, so therefore, it is necessary to be careful interpreting the results.

Summary Table for Rows, Columns
Row exclusion: Inspection

Num. Missing	0
DF	46
Chi Square	33.951
Chi Square P-Value	.9058
G-Squared	•
G-Squared P-Value	•
Contingency Coef.	.494
Cramer's V	.402

Figure 30. Chi-2 results for PG document.

Summary Table for Rows, Columns
Row exclusion: Inspection

Num. Missing	0
DF	46
Chi Square	41.676
Chi Square P-Value	.6538
G-Squared	•
G-Squared P-Value	•
Contingency Coef.	.535
Cramer's V	.448

Figure 31. Chi-2 results for ATM document.

The results show that there is no significant difference between the different perspectives. The results for the PG document show that it is almost the same defects that is found by the perspectives. The ATM differs a little bit, but still there is no significance.

To investigate how different (or, in this case, similar) the perspectives are, a Pearson correlation analysis was performed. The results indicate that there is a significant positive correlation between the perspectives, i.e., if one perspective finds a defect, the other perspectives are likely to find it as well. The low correlation between the tester’s and designer’s perspectives on the ATM document is noticeable.

Correlation Analysis

	Correlation	P-Value	95% Lower	95% Upper
User (PG), Tester (PG)	.463	.0092	.123	.706
User (PG), Designer (PG)	.543	.0016	.228	.756
Tester (PG), Designer (PG)	.601	.0003	.307	.790

30 observations were used in this computation.

Figure 32. Correlation analysis for PG document

Correlation Analysis

	Correlation	P-Value	95% Lower	95% Upper
User (ATM), Tester (ATM)	.480	.0076	.138	.720
User (ATM), Designer (ATM)	.499	.0052	.162	.732
Tester (ATM), Designer (ATM)	.258	.1789	-.120	.570

29 observations were used in this computation.
One case was omitted due to missing values.

Figure 33. Correlation analysis for ATM document.

The boxplots indicated a difference between the documents. To investigate the more in-depth and ANOVA test was performed for the effort, efficiency, and rate.

ANOVA Table for Time

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Document	1	15824.033	15824.033	6.312	.0180	6.312	.681
Residual	28	70193.333	2506.905				

Means Table for Time
Effect: Document

	Count	Mean	Std. Dev.	Std. Err.
ATM	15	201.200	52.838	13.643
PG	15	155.267	47.137	12.171

Figure 34. ANOVA test for effort

ANOVA Table for Efficiency

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Document	1	1640340.833	1640340.833	1.413	.2446	1.413	.198
Residual	28	32506569.467	1160948.910				

Means Table for Efficiency
Effect: Document

	Count	Mean	Std. Dev.	Std. Err.
ATM	15	2474.467	856.098	221.044
PG	15	2942.133	1260.553	325.473

Figure 35. ANOVA test for efficiency

ANOVA Table for Rate

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Document	1	10678.533	10678.533	1.595	.2170	1.595	.218
Residual	28	187441.333	6694.333				

Means Table for Rate
Effect: Document

	Count	Mean	Std. Dev.	Std. Err.
ATM	15	271.133	91.961	23.744
PG	15	233.400	70.227	18.133

Figure 36. ANOVA test for detection rate

The results show that there is a difference between the time spent on the documents, but there is no significant difference between their efficiency or detection rate. Even though they spend more time, they have the same detection rate and efficiency. One problem might be that the subjects reviewing the PG document do not find as many defects.

Conclusions. The results from this study indicate that there is no significant difference between the three different perspectives. Both the Chi-2 and the correlation calculations confirm this, even though we must be careful interpreting the Chi-2 results. There is a significant difference between the two documents related to the effort spent to find defects, but there is no significant difference between the efficiency and the detection rate. The p-values vary around 0.23 and are not enough to be sure that there is not a difference. This should not be a very serious threat because we have not compared the documents against each other. Finally, the main result is that we would not recommend using PBR.